Smooth discrimination analysis *

Enno MAMMEN,
Institut für Angewandte Mathematik,
Ruprecht-Karls-Universität Heidelberg
Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany

Alexandre B. TSYBAKOV
Institut de Statistique, URA CNRS 1321, Université Paris VI,
4 Place Jussieu, F 75252 Paris, France

January 19, 1998

Abstract

Discriminant analysis for two data sets in $I\!R^d$ with probability densities f and g can be based on the estimation of the set $G = \{x : f(x) \geq g(x)\}$. We consider applications where it is appropriate to assume that the region G has a smooth boundary. In particular, this assumption makes sense if discriminant analysis is used as a data analytic tool. We discuss optimal rates for estimation of G.

1991 AMS: primary 62G05 , secondary 62G20
Keywords and phrases: discrimination analysis, minimax rates, Bayes risk
Short title: Smooth discrimination analysis

*This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse", Humboldt-Universität zu Berlin

1 Introduction

Assume that one observes two independent samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ of $\mathbb{R}^d$-valued i.i.d. observations with densities f or g, respectively. The densities f and g are unknown. An additional random variable Z is observed that is assumed to have density f or g (and to be independent of X and Y). We consider the discrimination problem to classify if Z comes from f or g. Discrimination problems were studied by many authors (see e.g. the recent books of Devroye, Györfi and Lugosi (1996) and Vapnik (1996) and the references cited therein).

A discrimination decision rule is defined by a set $G \subset \mathbb{R}^d$. We attribute Z to f if $Z \in G$ and to g otherwise. For a decision rule G the Bayes risk $R(G)$ [with prior probabilities 1/2] is:

$$R(G) = \frac{1}{2}\{\int_{G^c} f(x)\ dx + \int_G g(x)\ dx\},$$

where G^c is the complement of G. The Bayes risk is minimized by

$$G^* = \{x : f(x) \geq g(x)\}.$$

Denote $R^* = R(G^*) = \min_G R(G)$. Our approach can be generalized to other Bayes priors.

Remark that $R(G^*) = \frac{1}{2}\int \min\{f(x), g(x)\}\ dx$ and that

$$R(G) - R(G^*) = \frac{1}{2} d_{f,g}(G, G^*),$$

where

$$d_{f,g}(G, G') = \int_{G \triangle G'} |f - g|(x) dx$$

is a distance defined on Lebesgue measurable subsets of $\mathbb{R}^d$.

Since the densities f and g are assumed to be unknown, the Bayesian rule G^* is not available and one has to use empirical rules $\tilde{G}_{n,m}$, i.e. set valued functions based on the observations $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.

A standard way of assessing the quality of a decision rule $\tilde{G}_{n,m}$ is to estimate how fast $R(\tilde{G}_{n,m})$ converges to the minimal possible value R^*. The convergence $R(\tilde{G}_{n,m}) \to R^*$ [in probability, almost surely or in mean, respectively] was proved for various estimates $\tilde{G}_{n,m}$ of G^*. Moreover, certain bounds on the accuracy of $\tilde{G}_{n,m}$ were obtained for finite sample sizes. The book of Devroye, Györfi and Lugosi (1996) gives the most up-to-date survey of such results.

In this paper we study optimality of decision rules $\tilde{G}_{n,m}$:

- how fast can $R(\tilde{G}_{n,m})$ converge to R^*, under [smoothness] assumptions on G^*?

- which decision rule $\tilde{G}_{n,m}$ attains the optimal rate of convergence?

To our knowledge, the only previous study of this problem was that of Marron (1983). Under smoothness assumptions on the densities f and g he proved that the optimal rates of convergence in discrimination are the same as those of the mean integrated squared error in density estimation. As an error criterion, Marron (1983) used the integrated (over all prior probabilities p from 0 to 1) difference $R_p(\hat{G}_{n,m}) - R_p^*$, where $R_p(\cdot), R_p^*$ are the respective Bayes risks when the prior probabilities of f and g are p and $1-p$. Our approach is quite different. We do not suppose that f and g are smooth. Instead, we put conditions on the possible sets G^*. In particular, we consider the case where the sets G^* are smooth enough [more precisely, that the boundary of G^* is smooth]. This leads to different optimal decision rules and different optimal rates of convergence. In the setup of Marron (1983) plug-in rules $\{x : \hat{f}_n(x) > \hat{g}_m(x)\}$ show up as asymptotically optimal. Here $\hat{f}_n$ and $\hat{g}_m$ are properly chosen nonparametric estimators of f and g. Our decision rules are direct minimum contrast estimators of G^*. The intermediate density estimation step is avoided.

We consider nonparametric discrimination as a problem of set estimation. One of its specific features, as compared to other set estimation problems [see e.g. Korostelev and Tsybakov (1993), Rudemo and Stryhn (1994), Mammen and Tsybakov (1995), Polonik (1995), Tsybakov (1997)], is that the nonstandard distance $d_{f,g}$ is inherent for the definition of the risk. We show that, under assumptions on the ε-entropy of the class of possible sets G^*, the empirical risk minimization rules of Vapnik and Červonenkis (1974) type converge with optimal rate, in the distance $d_{f,g}$ and in the distance d_Δ [Lebesgue measure of symmetric difference, see (1) below]. The rate of convergence depends on the smoothness [ε-entropy, respectively] of the class of possible sets G^* and on the local slope of the difference $f-g$ around the boundary $\{x : f(x) = g(x)\}$. It is interesting that in all the cases the convergence of Bayes risks turns out to be rather fast: the optimal rate is always better than $(n \wedge m)^{-1/2}$.

We prove upper and lower bounds on minimax risks of estimators of G^*. The proof of the upper bounds uses general results on minimum contrast estimates [cf. Birgé and Massart (1993) and van de Geer (1995)]. The proof of lower bounds is based on Assouad's lemma and is inspired by the approaches in Korostelev and Tsybakov (1993) and Tsybakov (1997).

2 The results

Here we introduce the empirical decision rules and formulate the results on optimal rates for discrimination. We start with some definitions.

Accuracy of the set estimates will be measured by the distances $d_{f,g}(G, G')$ and

$$d_\Delta(G, G') = \lambda(G \triangle G'), \tag{1}$$

where G and G' are (Lebesgue measurable) sets in $\mathbb{R}^d$ and where λ denotes the Lebesgue measure. We will consider the estimation of $G_K = G^* \cap K$, rather than G^*, where K is a compact subset of $\mathbb{R}^d$.

A basic element of the model is the class $\mathcal{G}$ of possible "candidate" sets G. This class is assumed to be given. It imposes, in turn, the restrictions on a class $\mathcal{F}$ of possible pairs (f, g). The results are given in a minimax framework, over the class $\mathcal{F}$. For a specified class $\mathcal{G}$ of subsets of K and for positive constants c_1, c_2, η_0 and α the class $\mathcal{F}$ is defined as

$$\begin{aligned} \mathcal{F} \;=\; & \{(f,g) : f \text{ and } g \text{ are probability densities on } \mathbb{R}^d, \\ & \{x \in K : f(x) \geq g(x)\} \in \mathcal{G}, \\ & \frac{1}{c_1} \leq f(x), g(x) \leq c_1 \text{ for } x \in K, \\ & \lambda\{x \in K : |f(x) - g(x)| \leq \eta\} \leq c_2 \eta^\alpha \text{ for } 0 < \eta \leq \eta_0\}. \end{aligned} \tag{2}$$

This definition makes sense if the constants c_1, c_2, η_0 and α are such that the class $\mathcal{F}$ is not empty (for example, it means that c_1 is large enough). This is what we assume in the sequel, without deriving explicitly the restrictions on these parameters. Also, we assume for convenience that $0 < \eta_0 < 1/2$.

Consider now the following decision rule

$$\hat{G}_{n,m} = \arg\min_{G \in \mathcal{G}} R_{n,m}(G), \tag{3}$$

where

$$R_{n,m}(G) = \frac{1}{2n} \sum_{i=1}^{n} \mathbf{I}(X_i \in G^c) + \frac{1}{2m} \sum_{i=1}^{m} \mathbf{I}(Y_i \in G)$$

denotes the empirical risk. Here and below $\mathbf{I}$ is the indicator function and $G^c = K \backslash G$. Clearly, $R_{n,m}(G)$ is an unbiased estimator of $R(G)$.

Although the definition of the empirical decision rule $\hat{G}_{n,m}$ is similar to that of Vapnik and Červonenkis (1974) (see also Vapnik (1996) and the references therein), there is an important difference. We consider the empirical risk minimization over a nonparametric class of possible sets (in particular, over a smoothness class), rather than over a parametric collection of sets. This allows to approach asymptotically as close as possible the true set G^* and to treat the optimality issue.

Note also that the set estimation procedure (3) is closely related to the maximum likelihood density support estimators of Mammen and Tsybakov (1995) and to the excess mass estimators of density level sets studied by Hartigan (1987), Müller and Sawitzki (1991), Müller (1993) and Polonik (1995).

We study now the rate of convergence of $\hat{G}_{n,m}$ to G_K. This rate depends on the δ-entropy $H_B(\delta)$ [with bracketing] of the metric space $(\mathcal{G}, d_\triangle)$. For $\delta > 0$, the quantity $H_B(\delta) = H_B(\delta, \mathcal{G}, d_\triangle)$ is defined as the minimal number, such that $N_B(\delta) = \exp H_B(\delta)$ is an integer and such that there exist pairs $(U_j, V_j), j = 1, \ldots, N_B(\delta)$, of subsets of $\mathcal{G}$ satisfying

(i) $U_j \subset V_j$, for $j = 1, \ldots, N_B(\delta)$,

(ii) $d_\triangle(U_j, V_j) \leq \delta$, for $j = 1, \ldots, N_B(\delta)$,

(iii) for any $G \in \mathcal{G}$ there exists a $j \in \{1, \dots, N_B(\delta)\}$ such that $U_j \subset G \subset V_j$.

In the sequel we denote the probability measure and the expectation in case of underlying densities f and g by $\mathbf{P}_{f,g}$ or $\mathbf{E}_{f,g}$, respectively.

Theorem 1
For a class $\mathcal{G}$ of subsets of a compact set $K \subset \mathbb{R}^d$ and for positive constants c_1, c_2, η_0 and α define the class $\mathcal{F}$ of pairs of densities (f, g) according to (2). Suppose that there exist positive constants $\rho < 1$ and A such that

$$H_B(\delta, \mathcal{G}, d_\triangle) \le A\delta^{-\rho}, \tag{4}$$

for $\delta > 0$ small enough. Then, for all $p \ge 1$,

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathcal{F}} (n \wedge m)^{\alpha p/[2+\alpha+\rho\alpha]} \mathbf{E}_{f,g} d_\triangle^p(\hat{G}_{n,m}, G_K) < \infty, \tag{5}$$

$$\lim \sup_{n \wedge m \to \infty} \sup_{(f,g) \in \mathcal{F}} (n \wedge m)^{[1+\alpha]p/[2+\alpha+\rho\alpha]} \mathbf{E}_{f,g} d_{f,g}^p(\hat{G}_{n,m}, G_K) < \infty. \tag{6}$$

Here $n \wedge m$ denotes the minimum of n and m.

Theorem 1 allows to treat a number of interesting special cases. First, a rather general example where (4) holds is that of Dudley's classes $\mathcal{G}$ [Dudley (1974), see also Mammen and Tsybakov (1995)]. These classes contain sets (possibly, disconnected) with piecewise smooth boundaries. For these classes the estimators $\hat{G}_{n,m}$ are extremely difficult to compute, and it is worthwhile to replace the minimization over $\mathcal{G}$ in (3) by a minimization over a finite ε-net on $\mathcal{G}$ (cf. Mammen and Tsybakov (1995)).

Another example is given by the class $\mathcal{G}$ of convex subsets G of $K = [0,1]^2$. The bound (4) for this class holds with $\rho = 1/2$ [Dudley(1974)]. A computationally efficient algorithm for constructing $\hat{G}_{n,m}$ (which is in this case a convex set with piecewise linear boundary) is proposed by Müller (1995).

Finally, Theorem 1 covers the case where $\mathcal{G}$ is a class of boundary fragments with smooth boundaries (cf. Korostelev and Tsybakov (1993)). For this case, that we discuss in detail, we derive lower bounds on the minimax risks and show that the rates of Theorem 1 cannot be improved. We define now boundary fragments with Hölder continuous boundaries. For given $\gamma > 0$ and $d \ge 2$ consider the functions $b(x_1, \dots, x_{d-1})$, $b : [0,1]^{d-1} \to [0,1]$ having continuous partial derivatives up to order l, where l is the maximal integer that is strictly less than γ. Denote $p_x(y)$ the Taylor polynomial of order l for $b(y), y \in [0,1]^{d-1}$, at a point $x \in [0,1]^{d-1}$. For a given $L > 0$, let $\Sigma(\gamma, L)$ be the class of functions b such that

$$|b(y) - p_x(y)| \le L|y - x|^\gamma \text{ for all } x, y \in [0,1]^{d-1},$$

where $|y|$ stands for the Euclidean norm of $y \in [0,1]^{d-1}$. A function b in $\Sigma(\gamma, L)$ defines a set $G_b = \{(x_1, \dots, x_d) \in [0,1]^d : 0 \le x_d \le b(x_1, \dots, x_{d-1})\}$. Such sets are called boundary fragments. Define the class

$$\mathcal{G}_{frag} = \{G_b : b \in \Sigma(\gamma, L)\}. \tag{7}$$

It is well known [see e.g. Dudley (1974)] that the δ-entropy with bracketing of this class of sets satisfies

$$H_B(\delta, \mathcal{G}_{frag}, d_\triangle) \le A\delta^{-[d-1]/\gamma}, \tag{8}$$

for some $A > 0$ and all $\delta > 0$ small enough. Thus, (4) is satisfied with $\rho = (d-1)/\gamma$.

For given positive constants $\gamma, L, c_1, c_2, \eta_0$, and α define the class $\mathcal{F} = \mathcal{F}_{frag}$ of pairs (f, g) of probability densities satisfying (2) with $\mathcal{G} = \mathcal{G}_{frag}$.

The next theorem states that in this particular model no better rates can be achieved than the rates given in the upper bounds of Theorem 1, with $\rho = (d-1)/\gamma$.

Theorem 2
Let $K = [0,1]^d$ and let $\mathcal{F} = \mathcal{F}_{frag}$ be as in (2) with $\mathcal{G} = \mathcal{G}_{frag}$. Then

$$\lim \inf_{n\wedge m\to\infty} \inf_{\tilde G_{n,m}} \sup_{(f,g)\in\mathcal{F}_{frag}} (n\wedge m)^{\alpha\gamma p/[(2+\alpha)\gamma+\alpha(d-1)]} \mathbf{E}_{f,g} d_\triangle^p(\tilde G_{n,m}, G_K) > 0, \tag{9}$$

$$\lim \inf_{n\wedge m\to\infty} \inf_{\tilde G_{n,m}} \sup_{(f,g)\in\mathcal{F}_{frag}} (n\wedge m)^{[1+\alpha]\gamma p/[(2+\alpha)\gamma+\alpha(d-1)]} \mathbf{E}_{f,g} d_{f,g}^p(\tilde G_{n,m}, G_K) > 0, \tag{10}$$

for every $p \ge 1$.

Theorems 1 and 2, together with (8), show that the rates of convergence $(n\wedge m)^{-\alpha\gamma/[(2+\alpha)\gamma+\alpha(d-1)]}$ and $(n\wedge m)^{-(1+\alpha)\gamma/[(2+\alpha)\gamma+\alpha(d-1)]}$ are optimal in the minimax sense on the class $\mathcal{F}_{frag}$ for the distances $d_\triangle$ and $d_{f,g}$ respectively, whenever $\gamma > d-1$. Note that for the distance $d_\triangle$ the rate is exactly the same as the optimal rate in the problem of level sets estimation [cf. Tsybakov (1997)].

The case, where $p = 1$ and the distance is $d_{f,g}$, makes a particular interest for the discrimination setup. In this case we get the following corollary on the asymptotic behaviour of Bayes risks. It states that the Bayes risk of the decision rule $\hat G_{n,m}$ converges with optimal rate to the minimal Bayes risk $R(G_K)$.

Corollary 1
Let $\gamma > d-1$. Then

$$\frac{\sup_{(f,g)\in\mathcal{F}_{frag}} [R(\hat G_{n,m}) - R(G_K)]}{\inf_{\tilde G_{n,m}} \sup_{(f,g)\in\mathcal{F}_{frag}} [R(\tilde G_{n,m}) - R(G_K)]} = \rho_{n,m},$$

where $\rho_{n,m}$ is a sequence with $\rho_{n,m} = O(1)$ and $\rho_{n,m}^{-1} = O(1)$, as $n\wedge m \to \infty$. Moreover,

$$\inf_{\tilde G_{n,m}} \sup_{(f,g)\in\mathcal{F}_{frag}} [R(\tilde G_{n,m}) - R(G_K)] = \tilde\rho_{n,m}(n\wedge m)^{-(1+\alpha)\gamma/[(2+\alpha)\gamma+\alpha(d-1)]},$$

where $\tilde\rho_{n,m}$ is a sequence with $\tilde\rho_{n,m} = O(1)$ and $\tilde\rho_{n,m}^{-1} = O(1)$, as $n\wedge m \to \infty$.

It is interesting that the rate of convergence of Bayes risks is rather fast. It is always faster than the "parametric" rate $(n \wedge m)^{-1/2}$, since $(1+\alpha)\gamma/[(2+\alpha)\gamma + \alpha(d-1)] > 1/2$, whenever $\gamma > d-1$. We end this section by some remarks on generalizations and extensions of the results.

Remark 1. Corollary 1 remains valid if one replaces $\mathcal{F}_{frag}$ by $\mathcal{F}_{Dudley}$ where $\mathcal{F}_{Dudley}$ is the class $\mathcal{F}$ defined as in (2), with $\mathcal{G}$ being a Dudley class. This follows immediately from the fact that the Dudley class of sets with smoothness γ contains the class of boundary fragments of the same smoothness. Thus, the empirical rules (3) attain the optimal Bayes risk rates on the Dudley classes as well.

Remark 2. Theorem 2 can be easily extended to boundary fragments with convex or monotone boundaries $b(\cdot)$, when $d = 2$. For the case of convex $b(\cdot)$, one should set $\gamma = 2$ and choose g_0 in the proof of Theorem 2 to have a parabolic level profile rather than a constant one (cf. the proof of Theorem 5.2 in Mammen and Tsybakov (1995)). Together with Theorem 1, this shows the rate optimality of the rule (3) for the case where

$$\mathcal{G} = \mathcal{G}_{conv} = \{\text{all closed convex subsets of } [0,1]^2\}.$$

The corresponding optimal rates are $(n \wedge m)^{-2\alpha/(4+\alpha)}$ and $(n \wedge m)^{-2(1+\alpha)/(4+\alpha)}$ for the distances $d_\triangle$ and $d_{f,g}$ respectively.

If $\mathcal{G} = \mathcal{G}_{mon} = \{G_b \subset [0,1]^2 : b(\cdot) \text{ is monotone nondecreasing}\}$, then Theorem 2 remains valid with $\gamma = 1$. This easily follows if one performs the proof of Theorem 2 with a density g_0 having a linear level profile, rather than a constant one. However, in the case of $\mathcal{G} = \mathcal{G}_{mon}$ Theorem 1 does not apply, since $\rho = 1$. A possible way to extend Theorem 1 to this case might be to consider the empirical rule on a sieve rather than on the whole class $\mathcal{G}_{mon}$, and to apply the technique of Barron, Birgé and Massart (1995).

We mention briefly some other straightforward generalizations. Analogous results hold for the choice of Bayes prior probabilities p and $1-p$, with $p \neq \frac{1}{2}$; then the set G should be defined as $\{x : pf(x) \geq (1-p)g(x)\}$. Furthermore, models with random sample sizes [see Vapnik (1996), Devroye, Györfi and Lugosi (1996)] can be easily covered. Another generalization concerns models with more than two populations.

3 Proofs

Proof of Theorem 1.

We use a result of van de Geer (1995) that we state, for convenience, as a lemma.

For a probability measure P, consider a class $\mathcal{H}$ of uniformly bounded functions h in $L_2(P)$. Suppose that the δ-entropy with bracketing $H_B(\delta, \mathcal{H}, L_2(P))$ satisfies, for some $0 < \nu < 2$ and $A > 0$, the inequality

$$H_B(\delta, \mathcal{H}, L_2(P)) \leq A\delta^{-\nu} \tag{11}$$

for all $\delta > 0$ small enough. Let h_0 be a fixed element in $\mathcal{H}$.

Lemma 1 *(Van de Geer (1995), Lemma 2.3)*
Let (11) be satisfied. Then there exist constants $D_1 > 0, D_2 > 0$ such that for a sequence of i.i.d. random variables $Z_1, \ldots, Z_n$ with distribution P it holds that

$$P\Big(\sup_{h \in \mathcal{H}: \|h-h_0\| \geq n^{-\frac{1}{2+\nu}}} \frac{|\frac{1}{\sqrt{n}} \sum_{i=1}^n \{(h-h_0)(Z_i) - \mathbf{E}(h-h_0)(Z_i)\}|}{\|h-h_0\|^{1-\frac{\nu}{2}}} > D_1 x \Big) \leq D_2 e^{-x} \quad (12)$$

for $x \geq 1$. Here $\|\cdot\|$ denotes the $L_2(P)$-norm.

W.l.o.g. assume in the sequel that $n \leq m$. For a given set G denote: $h_1(x) = \mathbf{I}(x \in G^c)$, $h_{1,0}(x) = \mathbf{I}(x \in G_K^c)$, $h_2(y) = \mathbf{I}(y \in G)$, $h_{2,0}(y) = \mathbf{I}(y \in G_K)$. Note that

$$R_{n,m}(G) - R_{n,m}(G_K) = \frac{1}{2n} \sum_{i=1}^n (h_1 - h_{1,0})(X_i) + \frac{1}{2m} \sum_{i=1}^m (h_2 - h_{2,0})(Y_i). \quad (13)$$

Clearly,

$$\mathbf{E}_{f,g}(R_{n,m}(G) - R_{n,m}(G_K)) = \frac{1}{2} d_{f,g}(G_K, G). \quad (14)$$

Observe also that

$$c_1^{-1} d_\triangle(G_K, G) \leq \|h_1 - h_{1,0}\|_f^2 = \int\limits_{G_K \triangle G} f(x)\, dx \leq c_1 d_\triangle(G_K, G) \quad (15)$$

and

$$c_1^{-1} d_\triangle(G_K, G) \leq \|h_2 - h_{2,0}\|_g^2 = \int\limits_{G_K \triangle G} g(x)\, dx \leq c_1 d_\triangle(G_K, G), \quad (16)$$

where $\|h\|_f^2 = \int h^2(x) f(x) dx$. Furthermore, (4), (15) and (16) entail that (11) holds with $\nu = 2\rho$.

Consider the random variable

$$V_{n,m} = \sqrt{n} \frac{R_{n,m}(G_K) - R_{n,m}(\hat{G}_{n,m}) + d_{f,g}(\hat{G}_{n,m}, G_K)/2}{d_\triangle^{(1-\rho)/2}(\hat{G}_{n,m}, G_K)}.$$

By definition of $\hat{G}_{n,m}$ we have $R_{n,m}(\hat{G}_{n,m}) \leq R_{n,m}(G_K)$. This implies

$$\sqrt{n} \frac{d_{f,g}(\hat{G}_{n,m}, G_K)}{d_\triangle^{(1-\rho)/2}(\hat{G}_{n,m}, G_K)} \leq V_{n,m}. \quad (17)$$

We consider now the event

$$E = \{d_\triangle(G_K, \hat{G}_{n,m}) > c_1 n^{-2/(2+\nu)}\}.$$

Taking into account (13) and (14), we obtain that, if E holds,

$$V_{n,m} \leq \frac{\sqrt{n}|\frac{1}{2n}\sum_{i=1}^{n}\{(h_1-h_{1,0})(X_i)-\mathbf{E}(h_1-h_{1,0})(X_i)\}|}{d_{\triangle}^{(1-\rho)/2}(\hat{G}_{n,m},G_K)}$$

$$+\frac{\sqrt{n}|\frac{1}{2m}\sum_{i=1}^{m}\{(h_2-h_{2,0})(Y_i)-\mathbf{E}(h_2-h_{2,0})(Y_i)\}|}{d_{\triangle}^{(1-\rho)/2}(\hat{G}_{n,m},G_K)}$$

$$\leq \sup_{G\in\mathcal{G}:d_{\triangle}(G,G_K)\geq c_1 n^{-\frac{2}{2+\nu}}} \frac{\sqrt{n}|\frac{1}{2n}\sum_{i=1}^{n}\{(h_1-h_{1,0})(X_i)-\mathbf{E}(h_1-h_{1,0})(X_i)\}|}{d_{\triangle}^{(1-\rho)/2}(G,G_K)}$$

$$+ \sup_{G\in\mathcal{G}:d_{\triangle}(G,G_K)\geq c_1 n^{-\frac{2}{2+\nu}}} \frac{\sqrt{n}|\frac{1}{2m}\sum_{i=1}^{m}\{(h_2-h_{2,0})(Y_i)-\mathbf{E}(h_2-h_{2,0})(Y_i)\}|}{d_{\triangle}^{(1-\rho)/2}(G,G_K)}$$

$$\leq \sup_{h\in\mathcal{H}:\|h-h_{1,0}\|_f\geq n^{-\frac{1}{2+\nu}}} \frac{\sqrt{n}|\frac{1}{2n}\sum_{i=1}^{n}\{(h-h_{1,0})(X_i)-\mathbf{E}(h-h_{1,0})(X_i)\}|}{c_1^{(\rho-1)/2}\|h-h_{1,0}\|_f^{1-\rho}}$$

$$+ \sup_{h\in\mathcal{H}:\|h-h_{2,0}\|_g\geq m^{-\frac{1}{2+\nu}}} \frac{\sqrt{m}|\frac{1}{2m}\sum_{i=1}^{m}\{(h-h_{2,0})(Y_i)-\mathbf{E}(h-h_{2,0})(Y_i)\}|}{c_1^{(\rho-1)/2}\|h-h_{2,0}\|_g^{1-\rho}},$$

where we used (15) and (16) to get the last inequality.

Thus, Lemma 1 implies

$$\lim \sup_{n\wedge m\to\infty} \mathbf{E}\left[\exp(t|V_{n,m}|)\mathbf{I}(E)\right] < \infty \tag{18}$$

for $t>0$ small enough.

We use now the following lemma.

Lemma 2

There exists a constant $c(\alpha)$ depending on α such that for every Lebesgue measurable subset G of K and for $(f,g)\in\mathcal{F}$,

$$c(\alpha)\; d_{\triangle}^{(1+\alpha)/\alpha}(G,G_K) \leq d_{f,g}(G,G_K) \leq 2c_1 d_{\triangle}(G,G_K).$$

On E^c we have $d_{\triangle}(G_K,\hat{G}_{n,m}) \leq c_1 n^{-1/(1+\rho)}$ and, because of the second inequality of Lemma 2, $d_{f,g}(G_K,\hat{G}_{n,m}) \leq 2c_1^2 n^{-1/(1+\rho)}$. Since

$$n^{-1/(1+\rho)} = o(n^{-[1+\alpha]/[2+\alpha+\rho\alpha]})$$

and

$$n^{-1/(1+\rho)} = o(n^{-\alpha/[2+\alpha+\rho\alpha]})$$

it suffices to consider the event E.

The first inequality of Lemma 2 and (17) imply

$$d_\triangle(G_K, \hat{G}_{n,m}) \le (V_{m,n}/c(\alpha))^{2\alpha/[2+\alpha+\rho\alpha]} n^{-\alpha/[2+\alpha+\rho\alpha]}. \tag{19}$$

Together with (18) this shows (5). Equation (6) can be proved by plugging (19) into (17). □

Proof of Lemma 2.

The second inequality of the lemma is trivial. To prove the first inequality note that the condition $\lambda(|f-g| < \eta) \le c_2\eta^\alpha, 0 < \eta \le \eta_0$ implies $\lambda(|f-g| < \eta) \le \tilde{c}_2\eta^\alpha, 0 < \eta \le 2c_1$, where $\tilde{c}_2 > 0$ depends only on c_1, c_2, η_0 and α. Hence, with an appropriate choice of η and $c(\alpha)$, we have

$$\begin{aligned} d_{f,g}(G, G_K) &\ge \int\limits_{G\triangle G_K} |f-g|\mathbf{I}(|f-g| \ge \eta) \\ &\ge \eta[\lambda(G\triangle G_K) - \lambda(|f-g| < \eta)] \\ &\ge \eta d_\triangle(G, G_K) - \tilde{c}_2\eta^{1+\alpha} \\ &\ge c(\alpha) d_\triangle^{(1+\alpha)/\alpha}(G, G_K). \end{aligned} \tag{20}$$

□

Proof of Theorem 2.

Suppose w.l.o.g. that $n \le m$. We consider the subset of $\mathcal{F}_{frag}$ that contains all pairs (f, g_0), where g_0 is a fixed density on K and f belongs to a finite class of densities $\mathcal{F}_1$ that will be defined below. Then

$$\begin{aligned} &\sup_{(f,g)\in\mathcal{F}_{frag}} \mathbf{E}_{f,g} d_\triangle^p(\tilde{G}_{n,m}, G) \\ &\ge \sup_{(f,g_0): f\in\mathcal{F}_1} \mathbf{E}_{f,g_0} d_\triangle^p(\tilde{G}_{n,m}, G) \\ &\ge \mathbf{E}_{g_0}\Big[\frac{1}{\#\mathcal{F}_1}\sum_{f\in\mathcal{F}_1} \mathbf{E}_f\{d_\triangle^p(\tilde{G}_{n,m}, G)|Y_1, \dots, Y_m\}\Big], \end{aligned} \tag{21}$$

where $\mathbf{E}_f$ and $\mathbf{E}_{g_0}$ denote the expectations w.r.t. the distributions of $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_m)$ when the underlying densities are f and g_0. Here and later $\#\mathcal{F}_1$ denotes the number of elements of $\mathcal{F}_1$.

For simplicity, we give the proof only for the case $d = 2$ (extension to higher dimensions is straightforward). For this case, it suffices to bound the term in squared brackets in (21) from below by $cn^{-\alpha\gamma p/[(2+\alpha)\gamma+1]}$, where $c > 0$ is a constant that does not depend on the sample $Y_1, \dots, Y_m$. This would prove (9). The lower bound (10) follows from (9) and Lemma 2. Furthermore, it suffices to consider the case $p = 1$, since it implies the result for $p \ge 1$ by application of the Hölder inequality. Hence for the proof of the theorem it suffices to show that for any estimator $\tilde{G}_{n,m}$ and any

n, m large enough

$$n^{\frac{\alpha\gamma}{(2+\alpha)\gamma+1}} \frac{1}{\#\mathcal{F}_1} \sum_{f \in \mathcal{F}_1} \mathbf{E}_f\{d_\Delta(\tilde{G}_{n,m}, G)|Y_1, \ldots, Y_m\} \geq c \qquad \text{(a.s.)}, \tag{22}$$

where $c > 0$ does not depend on n, m [and $Y_1, \ldots, Y_m$].

Before we come to the proof of (22) we define g_0 and the class $\mathcal{F}_1$. For this purpose, let φ be an infinitely many times differentiable function on $\mathbb{R}^1$ with the following properties: $\varphi(t) = 0$ for $|t| \geq 1$, $\varphi(t) \geq 0$ for all t, $\max_t \varphi(t) = 1$ and $\varphi(0) = 1$. For a fixed integer $M \geq 2$ and a constant τ with $0 < \tau < 1$ we define $b_1 = [\tau/c_2]^{1/\alpha} M^{-\gamma/\alpha}$. For $x = (x_1, x_2)$ in $K = [0,1]^2$ we put now

$$\begin{aligned} g_0(x) &= (1 + \eta_0 + b_1)\mathbf{I}\{0 < x_2 < \frac{1}{2}\} + \mathbf{I}\{\frac{1}{2} \leq x_2 < \frac{1}{2} + \tau M^{-\gamma}\} \\ &\quad + (1 - \eta_0 - b_2)\mathbf{I}\{\frac{1}{2} + \tau M^{-\gamma} \leq x_2 \leq 1\}, \end{aligned}$$

where $b_2 > 0$ is chosen such that $\int g_0(x)\, dx = 1$ and $0 < \eta_0 < 1/2$. W.l.o.g. we assume that c_1 [see (2)] is large enough, so that $c_1^{-1} < g_0(x) < c_1$ for all x in K. For $j = 1, \ldots, M$ we put

$$\varphi_j(t) = \tau M^{-\gamma} \varphi(M[t - \frac{2j-1}{M}]).$$

For vectors $\omega = (\omega_1, \ldots, \omega_M)$ of elements $\omega_j \in \{0, 1\}$ and for $t \in [0, 1]$ we define

$$b(t, \omega) = \frac{1}{2} + \sum_{j=1}^{M} \omega_j \varphi_j(t).$$

Put $\Omega = \{0, 1\}^M$. With this notation, define for $\omega \in \Omega$ and $x \in [0, 1]^2$

$$f_\omega(x) = g_0(x) + [\frac{b(x_1, \omega) - x_2}{c_2}]^{1/\alpha} \mathbf{I}\{\frac{1}{2} \leq x_2 \leq b(x_1, \omega)\} - b_3(\omega)\mathbf{I}\{\frac{1}{2} + \tau M^{-\gamma} < x_2 \leq 1\},$$

where $b_3(\omega) > 0$ is chosen such that $\int f_\omega(x)\, dx = 1$. Set now

$$\mathcal{F}_1 = \{f_\omega : \omega \in \Omega\}.$$

Let us first show that

$$(f_\omega, g_0) \in \mathcal{F}_{frag} \tag{23}$$

for all $\omega \in \Omega$.

Proof of (23). The equality $\int [g_0(x) - f_\omega(x)]\, dx = 0$ entails

$$\int_0^1 \int_{1/2}^{b(x_1,\omega)} [\frac{b(x_1, \omega) - x_2}{c_2}]^{1/\alpha}\, dx_2\, dx_1 = b_3(\omega)[\frac{1}{2} - \tau M^{-\gamma}].$$

This gives

$$
\begin{aligned}
b_3(\omega) &= \frac{1}{\frac{1}{2}-\tau M^{-\gamma}} \sum_{j=1}^{M} \omega_j \int_0^1 \int_{1/2}^{\frac{1}{2}+\varphi_j(t)} [\frac{\frac{1}{2}+\varphi_j(t)-u}{c_2}]^{1/\alpha}\, du\, dt \qquad (24)\\
&= \frac{c_2^{-1/\alpha}}{\frac{1}{2}-\tau M^{-\gamma}} \sum_{j=1}^{M} \omega_j \int_0^1 \int_0^{\varphi_j(t)} v^{1/\alpha}\, dv\, dt\\
&= \frac{c_2^{-1/\alpha}}{\frac{1}{2}-\tau M^{-\gamma}} \frac{\alpha}{\alpha+1} \sum_{j=1}^{M} \omega_j \int_0^1 \varphi_j(t)^{1+\alpha^{-1}}\, dt\\
&\le \frac{c_2^{-1/\alpha}}{\frac{1}{2}-\tau M^{-\gamma}} \frac{\alpha}{\alpha+1} M[\tau M^{-\gamma}]^{1+\alpha^{-1}} \int \varphi(Mt)^{1+\alpha^{-1}}\, dt\\
&= O(M^{-\gamma(1+\alpha^{-1})}).
\end{aligned}
$$

Hence $c_1^{-1} \le f_\omega \le c_1$ for c_1 and M large enough. Next, the set

$$\{x : f_\omega(x) \ge g_0(x)\} = \{x : 0 \le x_2 \le b(x_1, \omega)\}$$

belongs to $\mathcal{G}_{frag}$ since $b(\cdot, \omega) \in \Sigma(\gamma, L)$ for $\tau > 0$ small enough. To guarantee (23), it remains to show

$$\lambda\{x \in K : |f_\omega(x) - g_0(x)| \le \eta\} \le c_2 \eta^\alpha,$$

for $0 < \eta \le \eta_0$. But this follows from the fact that, for $0 < \eta \le \eta_0$,

$$
\begin{aligned}
&\{x \in K : |f_\omega(x) - g_0(x)| \le \eta\}\\
&= \{x \in K : 1/2 \le x_2 \le b(x_1, \omega), [\frac{b(x_1,\omega) - x_2}{c_2}]^{1/\alpha} \le \eta\}\\
&= \{x \in K : b(x_1, \omega) - c_2\eta^\alpha \le x_2 \le b(x_1, \omega)\}.
\end{aligned}
$$

Proof of (22). We use Assouad's lemma [see Bretagnolle and Huber (1979) and Assouad (1983)]. For our purposes it will be more convenient to apply the version of this lemma stated in Korostelev and Tsybakov (1993), which is adapted to the problem of estimation of sets.

For $j = 1, \ldots, M$ and for a vector $\omega = (\omega_1, \ldots, \omega_M)$, we write

$$
\begin{aligned}
\omega_{j0} &= (\omega_1, \ldots, \omega_{j-1}, 0, \omega_{j+1}, \ldots, \omega_M),\\
\omega_{j1} &= (\omega_1, \ldots, \omega_{j-1}, 1, \omega_{j+1}, \ldots, \omega_M).
\end{aligned}
$$

For $i = 0$ and $i = 1$ let P_{ji} be the probability measure corresponding to the distribution of $X_1, \ldots, X_n$ when the underlying density is $f_{\omega_{ji}}$. The expectation w.r.t. P_{ji} is denoted by $\mathbf{E}_{ji}$. Arguing as in (5.3) - (5.6) in Korostelev and Tsybakov (1993) we find that the sum

$$S = \frac{1}{\#\mathcal{F}_1} \sum_{f \in \mathcal{F}_1} \mathbf{E}_f\{d_\Delta(\tilde{G}_{n,m}, G) | Y_1, \ldots, Y_m\}$$

is bounded as follows:

$$\begin{aligned} S &\geq \frac{1}{2}\sum_{j=1}^{M}\lambda\{x:\frac{1}{2}\leq x_2\leq\frac{1}{2}+\varphi_j(x_1)\}\int\min\{dP_{j1},dP_{j0}\} \qquad (25)\\ &= \frac{1}{2}\sum_{j=1}^{M}\tau M^{-\gamma}\int\varphi(Mt)\ dt\ \int\min\{dP_{j1},dP_{j0}\}. \end{aligned}$$

Now,

$$\int\min\{dP_{j1},dP_{j0}\}\geq\frac{1}{2}[1-H^2(P^0,P^1)/2]^n,$$

where $H(\cdot,\cdot)$ denotes the Hellinger distance and P^0, P^1 denote the probability distributions of X_1 under the densities $f_{\omega_{10}}$ or $f_{\omega_{11}}$, respectively. We have

$$\begin{aligned} H^2(P^0,P^1) &= \int[\sqrt{f_{\omega_{10}}(x)}-\sqrt{f_{\omega_{11}}(x)}]^2\ dx \qquad (26)\\ &= \int_0^1\Big[\int_{\frac{1}{2}}^{\frac{1}{2}+\varphi_1(x_1)}\{1-\sqrt{1+(\frac{\frac{1}{2}+\varphi_1(x_1)-x_2}{c_2})^{1/\alpha}}\}^2\ dx_2\\ &\quad+\int_{\frac{1}{2}+\tau M^{-\gamma}}^{1}\{\sqrt{1-\eta_0-b_2-b_3(\omega_{10})}-\sqrt{1-\eta_0-b_2-b_3(\omega_{11})}\}^2\ dx_2\Big]\ dx_1\\ &\leq \int_0^1\Big[\int_0^{\varphi_1(x_1)}\{1-\sqrt{1+(\frac{v}{c_2})^{1/\alpha}}\}^2dv\Big]dx_1+\frac{1}{2}|b_3(\omega_{10})-b_3(\omega_{11})|^2. \end{aligned}$$

Here

$$\begin{aligned} &\int_0^1\int_0^{\varphi_1(x_1)}\{1-\sqrt{1+(\frac{v}{c_2})^{1/\alpha}}\}^2\ dv\ dx_1 \qquad (27)\\ &\leq \frac{1}{2}\int_0^1\int_0^{\varphi_1(x_1)}(\frac{v}{c_2})^{2/\alpha}\ dv\ dx_1\\ &= \frac{\alpha}{2(\alpha+2)}c_2^{-2/\alpha}\int_0^1[\varphi_1(x_1)]^{1+2\alpha^{-1}}\ dx_1\\ &= \frac{\alpha}{2(\alpha+2)}c_2^{-2/\alpha}[\tau M^{-\gamma}]^{1+2\alpha^{-1}}\int[\varphi(Mt)]^{1+2\alpha^{-1}}\ dt\\ &\leq C^*M^{-\gamma(1+2\alpha^{-1})-1}, \end{aligned}$$

where C^* depends only on α, c_2, τ and φ.

On the other hand, similarly to (24), one gets

$$\begin{aligned} |b_3(\omega_{10})-b_3(\omega_{11})| &\leq \frac{c_2^{-1/\alpha}}{\frac{1}{2}-\tau M^{-\gamma}}\frac{\alpha}{\alpha+1}\int_0^1[\varphi_j(t)]^{1+\alpha^{-1}}\ dt \qquad (28)\\ &= O(M^{-\gamma(1+\alpha^{-1})-1}). \end{aligned}$$

Combining (26)-(28), one gets

$$H^2(P^0,P^1)\leq C^*M^{-\gamma(1+2\alpha^{-1})-1}[1+o(1)].$$

Choose now M as the smallest integer that is larger or equal to $n^{\alpha/[(2+\alpha)\gamma+\alpha]}$. Then

$$H^2(P^0, P^1) \leq C^* n^{-1}[1 + o(1)].$$

This gives, with a constant $C_1^* > 0$,

$$\int \min\{dP_{j1}, dP_{j0}\} \geq \frac{1}{2}[1 - \frac{C^*}{2} n^{-1}\{1 + o(1)\}]^n \geq C_1^*$$

for all n large enough. This inequality and (25) yield

$$S \geq \frac{1}{2}\, C_1^* \tau M^{-\gamma} \int \varphi(t)dt \geq C_2^* n^{-\alpha\gamma/[(2+\alpha)\gamma+\alpha]},$$

for all n large enough. The constant $C_2^* > 0$ depends only on α, c_2, τ and φ. This finishes the proof of (22). Thus, the theorem is proved. □

References

[1] Assouad, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci., Paris,* Ser. 1 **296** 1021 - 1024.

[2] Barron, A., Birgé, L. and Massart, P. (1995). Risk bounds for model selection via penalization. *Technical Report* 95.54, Université Paris-Sud.

[3] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Prob. Theory Rel. Fields* **97** 113 - 150.

[4] Bretagnolle, J. and Huber, C. (1979) Estimation des densités: risque minimax. *Z. Warsch. verw. Geb.* **47** 119-137.

[5] Devroye, L. , Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* Springer, New York, Berlin, Heidelberg.

[6] Dudley, R.M. (1974) Metric entropy of some classes of sets with differentable boundaries. *J.Approx.Theory* **10** 227-236.

[7] Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267 - 270.

[8] Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction.* Lecture Notes in Statistics **82**, Springer, New York, Berlin, Heidelberg.

[9] Mammen, E. and Tsybakov, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502 - 524.

[10] Marron, J. S. (1983). Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Statist.* **11** 1142 - 1155.

[11] Müller, D. W. (1993). The excess mass approach in statistics. *Beiträge zur Statistik* **3** Universität Heidelberg.

[12] Müller, D. W. (1995). A backward-induction algorithm for computing the best convex contrast of two bivariate samples. *Beiträge zur Statistik* **29** Universität Heidelberg.

[13] Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738 - 746.

[14] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *Ann. Statist.* **23** 855 - 881.

[15] Rudemo, M. and Stryhn, H. (1994) Approximating the distributions of maximum likelihood contour estimates in two-region images. *Scand.J. of Statist.* **21** 41-56.

[16] Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25**, No.3.

[17] van de Geer, S. (1995). A maximal inequality for empirical processes. *Technical Report.*

[18] Vapnik, V. N. (1996). *The Nature of Statistical Learning Theory.* Springer, New York, Berlin, Heidelberg.

[19] Vapnik, V. N. and Červonenkis, A. Ja. (1974). *Theory of Pattern Recognition.* Nauka, Moscow (in Russian).